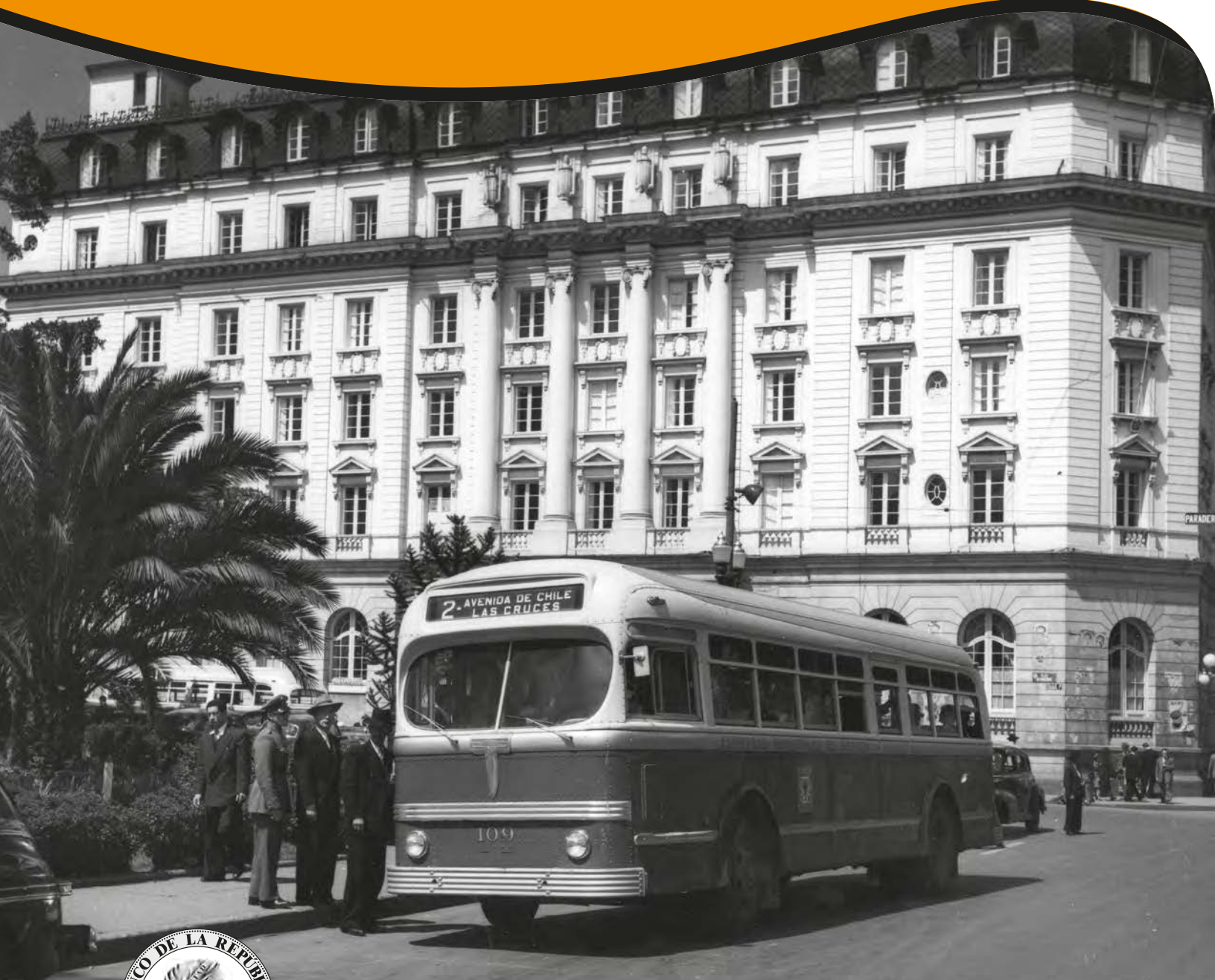


Descripción de las Minutas e Informes de Política Monetaria a partir de herramientas de Lingüística Computacional

Por: Daniela V. Guío-Martínez
Juan J. Ospina-Tejeiro
Germán A. Muñoz-Bravo
Julián A. Parra-Polanía

Núm. 1108
2020

Borradores de ECONOMÍA



Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia - Bogotá - Colombia

Descripción de las Minutas e Informes de Política Monetaria a partir de herramientas de Lingüística Computacional

Daniela V. Guío-Martínez*
dvguiom@unal.edu.co

Germán A. Muñoz-Bravo
gmunozbr@banrep.gov.co

Juan J. Ospina-Tejeiro
jospinte@banrep.gov.co

Julián A. Parra-Polanía
jparrapo@banrep.gov.co

Banco de la República

Las opiniones contenidas en el presente documento son responsabilidad exclusiva de los autores y no comprometen al Banco de la República ni a su Junta Directiva

RESUMEN

Con base en el uso de Latent Dirichlet Allocation, una herramienta de lingüística computacional cuya finalidad es develar los patrones temáticos subyacentes que agrupan las palabras de un conjunto de textos, analizamos dos tipos esenciales de documentos en la comunicación del Banco de la República, las minutas y los informes de política monetaria, para el periodo comprendido entre marzo de 2007 y diciembre de 2018. Encontramos que estos dos tipos de documentos giran primordialmente en torno a ocho temas, siendo el más importante (en promedio a través del tiempo) el que contiene términos principalmente relacionados con demanda interna y sectores económicos. Describimos tanto las similitudes como las diferencias que se observan, entre las minutas y los informes, en la participación de cada tema dentro de los documentos y en la evolución de esa participación en el tiempo.

Códigos JEL: E52, E58, C40

Palabras clave: Comunicación, Política Monetaria, Minería de Texto, LDA

* D. V. Guío-Martínez participó en el proyecto durante la práctica estudiantil en el Banco de la República. Los demás autores son empleados del Banco de la República de Colombia.

Description of the Minutes and Monetary Policy Reports using Computational Linguistics tools

Daniela V. Guío-Martínez
dvguiom@unal.edu.co

Germán A. Muñoz-Bravo
gmunozbr@banrep.gov.co

Juan J. Ospina-Tejeiro
jospinte@banrep.gov.co

Julián A. Parra-Polanía
jparrapo@banrep.gov.co

Banco de la República

<p>The opinions contained in this document are the sole responsibility of the authors and do not commit Banco de la República or its Board of Directors</p>

ABSTRACT

Based on the use of Latent Dirichlet Allocation, a computational linguistics tool whose purpose is to identify the underlying thematic patterns that group the words of a set of documents, we analyse two essential outlets in the Banco de la República's communication, the minutes and monetary policy reports, from March 2007 to December 2018. We find that these two outlets discuss primarily about eight topics, the most important (on average, over time) being the one that contains expressions mainly related to domestic demand and economic sectors. We describe both similarities and differences that are observed, between the minutes and the reports, in the participation of each topic within the documents and in the evolution of that participation over time.

JEL codes: E52, E58, C40

Keywords: Communication, Monetary Policy, Text Mining, LDA

* D. V. Guío-Martínez contributed to this project during a student internship at Banco de la República. The other authors are employees at Banco de la República de Colombia.

INTRODUCCIÓN

La comunicación hace parte fundamental del diseño y la ejecución de la política monetaria. En el proceso de toma de decisiones de los agentes económicos no solo resultan importantes las acciones de los encargados de la política económica, sino también la percepción acerca del propósito y la justificación de esas acciones y las expectativas sobre el curso de la economía y las acciones de política futuras.

La comunicación de los bancos centrales se realiza a través de distintos medios, entre ellos las conferencias de prensa, las minutas, los reportes o informes periódicos, y los discursos y entrevistas de los miembros del comité de política monetaria (o “la junta directiva” como se denomina en Colombia). A través de estos medios se entrega información de tipo numérico (v.g. un pronóstico numérico de crecimiento) y narrativo (v.g. una consideración cualitativa sobre las tasas de interés internacionales). Analizar la información de tipo numérico es un ejercicio más común en economía, mientras que en campos como la ciencia política o la sociología se hace un uso amplio del análisis de información narrativa.

Con el fin de ampliar las posibilidades de recolección y análisis de datos e incluir información de tipo narrativo, en los últimos años se ha aplicado con mayor frecuencia el procesamiento de texto¹ o de lenguaje natural² en el campo de la economía. En el caso particular de la política monetaria, el uso de estos métodos puede resultar útil tanto para los bancos centrales, con el fin de mejorar el análisis y la recolección de variables para el

¹ Este consiste en la aplicación de herramientas informáticas para el análisis de texto con el fin de recopilar una serie de patrones que ayudan a su interpretación.

² El lenguaje natural se define como la forma de comunicación cotidiana de los seres humanos (español, inglés, hindi etc).

proceso de toma de decisiones, como para los analistas financieros, por su interés en la extracción de la mayor cantidad de información a partir de las diferentes formas de comunicación de los bancos centrales.

En el presente documento se utilizan herramientas de lingüística computacional o, en otras palabras, de procesamiento de lenguaje natural mediante el uso de computadores para analizar dos formas de comunicación con el público muy importantes para el Banco de la República (BR): las minutas de las reuniones de política monetaria y los Informes de Política Monetaria –IPM- (titulados hasta el segundo trimestre de 2019 como Informes sobre Inflación). Las primeras corresponden a las reseñas de la deliberación hecha por los miembros de la Junta Directiva del BR (JDBR) durante las reuniones en las que se toman las decisiones sobre la tasa de política monetaria. Los segundos contienen la recopilación de información, los detalles técnicos del análisis y las perspectivas de política monetaria y de la economía, elaborados de forma independiente por el equipo técnico del BR y presentados a la JDBR como un insumo para su proceso de decisión.

De acuerdo con el manual elaborado por Bohlat et al. (2015), existen varias técnicas de procesamiento de texto usadas en el ámbito económico. Por un lado están los métodos booleanos³ y de diccionario, en los cuales se analizan los textos a partir de un conjunto predefinido de palabras clave. Por otro lado, hay técnicas de mayor sofisticación estadística en las que se supone que las palabras o términos en un documento están relacionadas por temas subyacentes en el mismo. El propósito de estos métodos es inferir o develar esos patrones temáticos que agrupan las palabras de un conjunto de documentos. En este enfoque se encuentra, entre otros, el método del que se hace uso en el presente trabajo:

³ Se refiere principalmente a búsquedas de expresiones usando operadores lógicos (AND, OR, NOT)

Latent Dirichlet Allocation (LDA), el cual se explica con mayor detalle en la siguiente sección.

Una característica sobresaliente y muy útil de LDA es que sus resultados no están restringidos a que cada término deba corresponder a un único tema. LDA se basa en un modelo (probabilístico) en el que cada palabra y cada documento está relacionado con varios temas, lo que además permite flexibilidad semántica (esto es, una misma palabra puede ser asignada a diferentes contextos, con diferentes significados). Cada documento puede verse como una distribución de probabilidad específica sobre los temas y cada tema es representado por una lista de palabras con sus correspondientes probabilidades de ocurrencia. Con base en esta herramienta, el presente trabajo analiza tanto minutas como los IPM e identifica las respectivas distribuciones de los tópicos a lo largo del periodo comprendido entre marzo de 2007 y diciembre de 2018.

El número de temas (y la distribución de palabras en cada uno de ellos) se determina de forma conjunta, esto es, se parte del supuesto de que las minutas y los informes se refieren exactamente a los mismos temas. Mediante el uso de un índice, que intenta medir la coherencia de los tópicos resultantes, se llega a que el modelo más apropiado es uno que distribuye las palabras (y los documentos) en ocho temas. Con base en las palabras más significativas para cada uno, se concluye que los temas están relacionados con política monetaria, demanda interna, balanza comercial, expectativas de inflación, inflación de alimentos, crédito, sector externo, y proyecciones macroeconómicas.

Por el contrario, la participación o distribución de los temas se estima por separado para cada documento. Con respecto a la evolución de esas participaciones en el tiempo se observan algunas diferencias importantes entre las minutas y los informes, como se muestra

en las gráficas presentadas en la segunda sección (“Resultados”). En particular, los resultados sugieren que la información narrativa contenida en las minutas corresponde a un análisis más prospectivo en comparación con el contenido narrativo de los IPM, lo que puede ser resultado tanto del hecho de que estos últimos regularmente incluyen un análisis retrospectivo (de la inflación y la actividad económica reciente) y que una parte importante de su análisis prospectivo se entrega en forma gráfica y/o numérica (v.g. fan charts de inflación y crecimiento económico).⁴

Por otro lado, los resultados obtenidos también muestran que existen algunas similitudes en la evolución de las distribuciones de temas. Por ejemplo, el tema más importante en promedio para el periodo analizado tanto para las minutas como para los IPM, es el relacionado con Demanda Interna y Sectores Económicos. Adicionalmente, tanto en minutas como en los informes se presenta una disminución de la participación del tema de inflación de alimentos y regulados y, en contraste, un aumento de la participación para el caso del tema de las proyecciones macroeconómicas.

Para el caso colombiano, existen algunos trabajos que han hecho uso de herramientas de análisis de texto para las comunicaciones del banco central. Taborda (2015) analiza las minutas de cinco bancos centrales (Brasil, Chile, Colombia, México y Perú). El autor examina en particular su longitud, vocabulario y lecturabilidad (es decir, qué tan fácil es comprender el texto) y la evolución de estas tres variables en el tiempo (2007-

⁴ Recientemente (octubre de 2019), el Banco de la República implementó algunos cambios que, por un lado, reducen el rezago de publicación, tanto de las minutas como de los IPM, con respecto a las correspondientes reuniones de Junta y, por otro lado, introducen algunas modificaciones en la estructura del IPM. Estas últimas implican, entre otras cosas, un mayor énfasis en el análisis prospectivo y de las proyecciones económicas.

2014, para el caso colombiano). Particularmente para Colombia el autor encuentra, entre otros resultados y para el periodo analizado, que las minutas tuvieron una tendencia creciente en su longitud, que el vocabulario sugiere que la inflación de alimentos ha sido especialmente importante (en comparación con los otros cuatro países), y que el nivel de lecturabilidad de las minutas se mantuvo relativamente constante en un nivel intermedio.

Arango et al. (2017), analizan los comunicados de prensa del Banco de la República para el periodo septiembre de 2004 a marzo de 2016. Los autores hacen uso de Latent Semantic Analysis (LSA), una herramienta similar a LDA en el sentido de intentar identificar la evolución temática en el conjunto de documentos, pero en la que los tópicos no son probabilísticos (esto es, en LSA las palabras y los documentos son asignados a los temas con probabilidad uno). Usando LSA, los autores identifican los temas relevantes y encuentran que de las minutas se infiere que la autoridad monetaria en Colombia le da mayor peso a la inflación que a cualquier otra variable. Luego, usando los temas identificados para la estimación de un VAR estructural (que incluye además las expectativas de inflación y un indicador de la situación económica), encuentran que las comunicaciones del Banco de la República pueden considerarse una herramienta de política monetaria: los comunicados de prensa tienen un efecto sobre las expectativas de inflación de los agentes en el mercado.

En Bancolombia (2019), el grupo de Investigaciones Económicas de esa entidad aplica diferentes métodos de análisis de texto a los comunicados y minutas (desde julio de 2007 hasta abril de 2019) con el propósito de reducir los errores de pronóstico sobre la tasa de política del Banco de la República. Específicamente, los autores hacen uso de las frecuencias de las palabras, y de los métodos LSA y LDA. A partir de estos métodos, entre

otros resultados, concluyen que la actividad productiva y el sector externo son los tópicos más relevantes en las discusiones de la JDDBR; pronostican que el próximo movimiento de tasas sería al alza, en 2020; y encuentran que incluir la información relacionada con los tópicos subyacentes en los comunicados y minutas del BR, les permite reducir en promedio un 24% el error de pronóstico sobre la tasa de política monetaria.

En adición a la información narrativa de las minutas y con base en la herramienta LDA, el presente trabajo analiza, como se mencionó anteriormente, los IPM. Hasta donde conocemos, estos últimos no han sido analizados con herramientas de lingüística computacional para el caso colombiano. De esta forma, se comparan dos formas de comunicación esenciales para el Banco de la República, uno elaborado a partir del proceso de deliberación de la Junta Directiva y otro del análisis realizado de forma independiente por el equipo técnico.

En adición a la introducción, este trabajo se divide en tres secciones. La primera de ellas describe las herramientas estadísticas usadas para el análisis del texto contenido en las minutas y los IPM. La segunda sección presenta los resultados obtenidos. La tercera sección concluye.

1. METODOLOGÍA

El aprendizaje automático, más conocido como *Machine Learning*, es una disciplina científica que consiste en crear (programar) sistemas que identifiquen patrones complejos de forma automática a partir de una serie de datos. Los sistemas de aprendizaje automático pueden clasificarse como supervisados y no supervisados (Géron, 2017), en donde los primeros hacen referencia a aquellos casos en que los datos son inicialmente clasificados

por seres humanos y las máquinas son “entrenadas” para seguir criterios similares de clasificación, y los segundos se refieren a los sistemas en los que las máquinas no se basan en ninguna clasificación predefinida por los humanos.

Diferentes técnicas se han desarrollado para ampliar el uso del aprendizaje automático, no solo a datos cuantitativos sino también al procesamiento del lenguaje natural, aplicando una serie de herramientas computacionales y estadísticas que cuantifican el texto (Bholat et al., 2015). Específicamente, el procesamiento de texto implica el uso de algoritmos de aprendizaje automático para ayudar a la comprensión y análisis de los documentos.

Dentro de la categoría del aprendizaje no supervisado se encuentran los métodos Latent Semantic Analysis (LSA) y Latent Dirichlet Allocation (LDA), entre otros. LSA representa un conjunto de documentos a través de una matriz de términos (o palabras), que captura la frecuencia de cada palabra en dicho conjunto, y a partir de la cual se buscan las combinaciones lineales que expliquen la mayor parte de la varianza de los términos a través de los documentos y las combinaciones lineales que expliquen la mayor parte de la varianza de los documentos a través de los términos. La hipótesis del modelo es que los componentes principales representan temas que son compartidos entre los documentos, mientras que los componentes descartados son temas de carácter idiosincrásico.

A diferencia de LSA, en donde cada término y cada documento son asignados a los temas con probabilidad uno, LDA supone que las palabras de un conjunto de documentos están relacionadas mediante temas, o tópicos no observados, de forma probabilística (por ejemplo, dado un tema en particular, una palabra específica haría parte de ese tema con cierta probabilidad). Más detalladamente, LDA descrito en Blei, Ng y Jordan (2003), es un

modelo probabilístico generativo en el que cada elemento, de una colección de datos discretos, se modela como una combinación finita sobre un conjunto subyacente de temas. LDA permite identificar los tópicos comunes presentes en el corpus (conjunto o colección de palabras) analizado, de esta forma, cualquier documento puede ser pensado como el resultado de un número específico de tópicos, cada uno con cierta probabilidad de ocurrencia y, a su vez, caracterizado por una distribución de probabilidad sobre las palabras.

Formalmente, para un conjunto de observaciones consistente en una serie de documentos de texto (y una cantidad de palabras por cada uno), la estructura subyacente (no observable) se refiere a la distribución de probabilidad sobre las palabras para un tópico k , denotada por β_k , la distribución de los tópicos para cada documento d , denotada por θ_d , y la asignación de los tópicos para cada palabra i en cada documento d , denotado por z_{di} .

Las distribuciones subyacentes, β , θ y z quedan condicionadas a la única variable observable del modelo, las palabras en los documentos (w), y se intentan inferir (de forma bayesiana) estimando la distribución de probabilidad (condicional) posterior mediante:

$$p(\mathbf{z}, \boldsymbol{\theta}, \boldsymbol{\beta} | \mathbf{w}) = \frac{p(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w})}{p(\mathbf{w})}$$

El cálculo de esta distribución posterior se convierte en un problema intratable, por lo que existen varias técnicas para solucionarlo de forma aproximada⁵. Estos métodos, sin embargo, pueden adolecer de algunos problemas, principalmente presentar restricciones computacionales importantes frente a conjuntos masivos de datos. Debido a esto, Hoffman,

⁵ Existen, por ejemplo, métodos variacionales que intentan encontrar un óptimo local a partir de una aproximación analítica a la distribución posterior o los métodos de muestreo (Monte Carlo) que usan en cambio una aproximación numérica a dicha distribución.

Blei y Bach (2010) desarrollan un algoritmo aproximado de inferencia que puede analizar una gran cantidad de documentos y que es utilizado en el presente documento⁶.

Finalmente, LDA proporciona una distribución conjunta sobre las variables subyacentes y las observables, y puede emplearse para obtener distribuciones posteriores de las primeras, dados los documentos. Así, una vez estimada la estructura subyacente del conjunto de documentos, se pueden estimar las distribuciones de tópicos por documento y por tanto la proporción en que se habla de un tópico en cada documento.

1.1. Preparación del corpus

Un paso importante para la cuantificación del texto es la selección de los documentos que harán parte del *corpus* de entrenamiento del modelo⁷. Para el caso del presente documento, el *corpus* está conformado por los IPM que publica el equipo técnico del Banco de la República y las minutas publicadas por la Junta Directiva de la misma entidad (JDDBR). El informe contiene varias secciones, de las cuales se tienen en cuenta las pertenecientes al contexto internacional, situación actual y perspectivas de crecimiento, desarrollos de la inflación y pronósticos. Se excluyen los recuadros y anexos, la introducción, los pies de página, texto en los cuadros y en los gráficos, así como las frases que se destacan ocasionalmente en los extremos laterales de las páginas ya que estas no hacen parte del cuerpo principal del texto.

⁶ Para el desarrollo del modelo se utilizó la librería *gensim* para Python. Para conocer más detalles véase: <https://radimrehurek.com/gensim/>.

⁷ Teniendo en cuenta la sección anterior, las distribuciones de probabilidad de los tópicos sobre las palabras se infieren automáticamente a partir de un *corpus* de entrenamiento, de ahí que el modelo estimado pueda aplicarse a otros documentos similares.

En el presente trabajo, se dispone de de 216 textos, 80 informes y 136 minutas, correspondientes al periodo comprendido entre el primer trimestre de 1999 y el tercer trimestre de 2018, para el caso de los primeros, y junio entre de 2007 y enero de 2019, para el caso de las segundas. Para llevar a cabo el entrenamiento, se sigue una estrategia similar a la usada por Hansen, McMachon & Tong (2018), quienes, aunque interesados directamente en los Reportes de inflación del Banco de Inglaterra, hacen la estimación a nivel de párrafos para explotar la variación de miles de ejemplos. Considerando, para efectos de entrenamiento, cada párrafo como un documento separado, se tienen en conjunto 17.833 documentos.

Los documentos en formatos binarios (v.g. PDF), que tienen disposiciones gráficas y diferentes áreas de texto que son excluidas del análisis, generalmente pasan por un proceso OCR⁸ para extraer solo el texto que se desea. Otra parte del proceso ocurre como resultado de que algunas palabras pueden estar mal escritas en los documentos, por lo que se debe realizar un paso de verificación adicional para cada archivo de texto.

Tokenización

Consiste en el proceso de segmentación automática de las frases, esto es, una vez se ha perfeccionado el corpus, cada documento pasa por un proceso de identificación de *tokens* individuales (términos/palabras) en el texto. Específicamente, los tokens son componentes textuales independientes y mínimos que tienen una sintaxis y una semántica definida (Sarkar, 2016), por ejemplo *inflacion* o *economías_emergentes*.

⁸ OCR son las siglas de Optical Character Recognition, y se refiere al reconocimiento de texto en imágenes. El software ABBYY FineReader, usado en este proceso específico, permite al usuario especificar el orden de las áreas de texto a extraer.

Stopwords

Otro proceso de preparación del corpus consiste en conformar y remover una lista de palabras que no aportan al sentido del texto por sí solas (*stopwords*) como por ejemplo los artículos, pronombres y preposiciones.

A pesar de que existen varias herramientas que proveen un conjunto de stopwords estandarizadas, es necesario personalizarlas teniendo en cuenta que los diferentes dominios temáticos utilizan palabras y frecuencias diferentes. Una lista adecuada de stopwords puede identificarse luego de varias iteraciones de entrenamiento del modelo, ya que aparecerán como palabras con una alta probabilidad dentro de los tópicos resultantes pero sin que agreguen valor a la comprensión del texto. En total, del corpus se eliminaron 828 stopwords⁹.

Part-of-Speech

Se refiere a categorías léxicas a las que se asignan las palabras en función de su contexto sintáctico y su función (Sarkar, 2016). Específicamente, el procedimiento computacional clasifica las palabras en categorías como *sustantivo, verbo, adjetivo o adverbio, numérico, puntuación*, entre otras. Una ventaja de este procedimiento es que permite excluir los *tokens* que son asociados a categorías que el investigador puede considerar como innecesarios a la hora de interpretar los tópicos y que pueden sesgar o generar ruido en el resultado. Para el análisis del presente documento se excluyeron los términos identificados como numéricos, puntuación y símbolo.

Bigramas y n-gramas

⁹ En adición a las palabras que tiene el programa por defecto (preposiciones, artículos...), se incluyeron palabras como “comportamiento”, “estadístico”, “cierre”, “mostrando”, “resultados”, “orden”, entre otras.

Un n -grama es una secuencia de n palabras que ocurren juntas con relativa frecuencia. Casos especiales de este modelo son los llamados bigramas o trigramas¹⁰. Los n -gramas se incluyen en el análisis como *tokens* individuales, ya que estos tienen mayor sentido para el documento cuando se tratan como único elemento y, de esta forma, se agrega valor al proceso de analizar diferencias entre los tópicos.

En resumen, los datos que van a servir de insumo para el modelo consisten en una lista de *tokens* para cada documento, de los cuales se excluyen los no alfabéticos, los denominados *stop-words* y aquellas categorías léxicas innecesarias para la interpretación del texto, como se explicó anteriormente (*part-of-speech*). Cada *token* se representa en letras minúsculas y se identifican los n -gramas relevantes. Por último, se filtran los tokens que aparecen en más del 50% de los documentos y los que aparecen en un único documento. Esto ayuda a reducir la carga computacional, dejando por fuera aquellos términos que no aportan significativamente a la explicación de la variabilidad entre documentos.

El modelo LDA está basado en una representación matricial del texto, también conocida como “bolsa de palabras” (bag of words), que se construye a partir de la frecuencia de los términos dentro de los documentos. Cada línea tiene dos números, el primero corresponde a la identificación única del *token* y el segundo a la frecuencia del mismo en el documento. Lo anterior a su vez implica que el algoritmo de entrenamiento comienza haciendo uso de estas frecuencias, independientemente del contexto específico en el que se utiliza cada *token* y, en el proceso de optimización y búsqueda de las

¹⁰ Ejemplos de estos n -gramas incluyen nombres de países u organizaciones e indicadores económicos, entre otras: “estados_unidos”, “economías_emergentes”, “economías_avanzadas”, “términos_intercambio” etc.

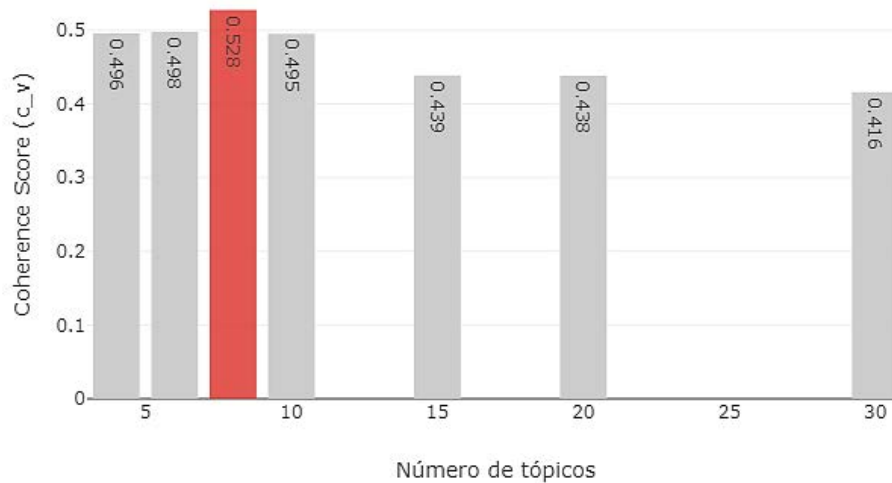
distribuciones posteriores, termina por asociar las palabras, los temas y los documentos, de manera que permitan explicar una buena proporción de la variabilidad entre estos últimos.

1.2. Selección del número de tópicos

El número de tópicos k es el parámetro principal a seleccionar cuando se entrena un modelo LDA; sin embargo, no hay un consenso sobre cómo determinar *a priori* el número de tópicos apropiado para un corpus específico. La elección de un número bajo dará lugar a tópicos muy generales mientras que un número alto, a temas muy específicos.

Técnicamente, la elección de un número adecuado de tópicos puede hacerse con base en la bondad de ajuste de los modelos, intentando que la interpretación de los resultados sea más fácil. Existen varias medidas para calcular el número óptimo de tópicos basadas en la hipótesis que afirma que las palabras con significado cercano tienden a ocurrir conjuntamente dentro de un contexto similar (Syed & Spruit, 2017). Una de estas medidas es el índice de coherencia (*Coherence Score*) descrito en Röder, Both y Hinneburg (2015). Dado un número específico de tópicos escogido por el investigador, este índice segmenta las palabras en subconjuntos, estima la distancia relativa entre palabras dentro de cada tópico y genera un puntaje agregado de coherencia (correspondiente a ese número de tópicos específico). De esta forma se pueden comparar los índices de coherencia calculados para varios números de tópicos, buscando escoger el que arroje el valor más alto.

Los números de tópicos con los que se entrenó el modelo fueron $k = 4,6,8,10,15,20,30$. Como lo muestra la Gráfica 1, el índice de coherencia más alto se encontró para el modelo con 8 tópicos.



Gráfica 1. Índice de Coherencia (*Coherence Score*) para cada número de tópicos.

Elaboración propia.

1.3. Algunas herramientas visuales para el análisis de los resultados

Después de seleccionar el modelo para ocho tópicos, se examinan con cuidado los resultados arrojados por ese modelo, esto es, a qué se refiere cada uno de esos temas, su importancia y la relación que tienen entre sí. Una forma de hacerlo es a través de la observación gráfica de los resultados, que puede tornarse difícil debido a la alta dimensionalidad del modelo. Para facilitar esta tarea, existe una herramienta de visualización interactiva cuya idea básica es inferir tanto la frecuencia del término como su exclusividad¹¹. Particularmente, se hace uso del sistema LDAvis, descrito en Sievert y Shirley (2014), el cual tiene dos características importantes: permite seleccionar un tema y

¹¹ En este contexto, exclusividad hace referencia a si la ocurrencia del término o palabra se limita a muy pocos tópicos.

mostrar los términos más relevantes para el mismo, y permite seleccionar un término para mostrar su distribución condicional sobre los temas.

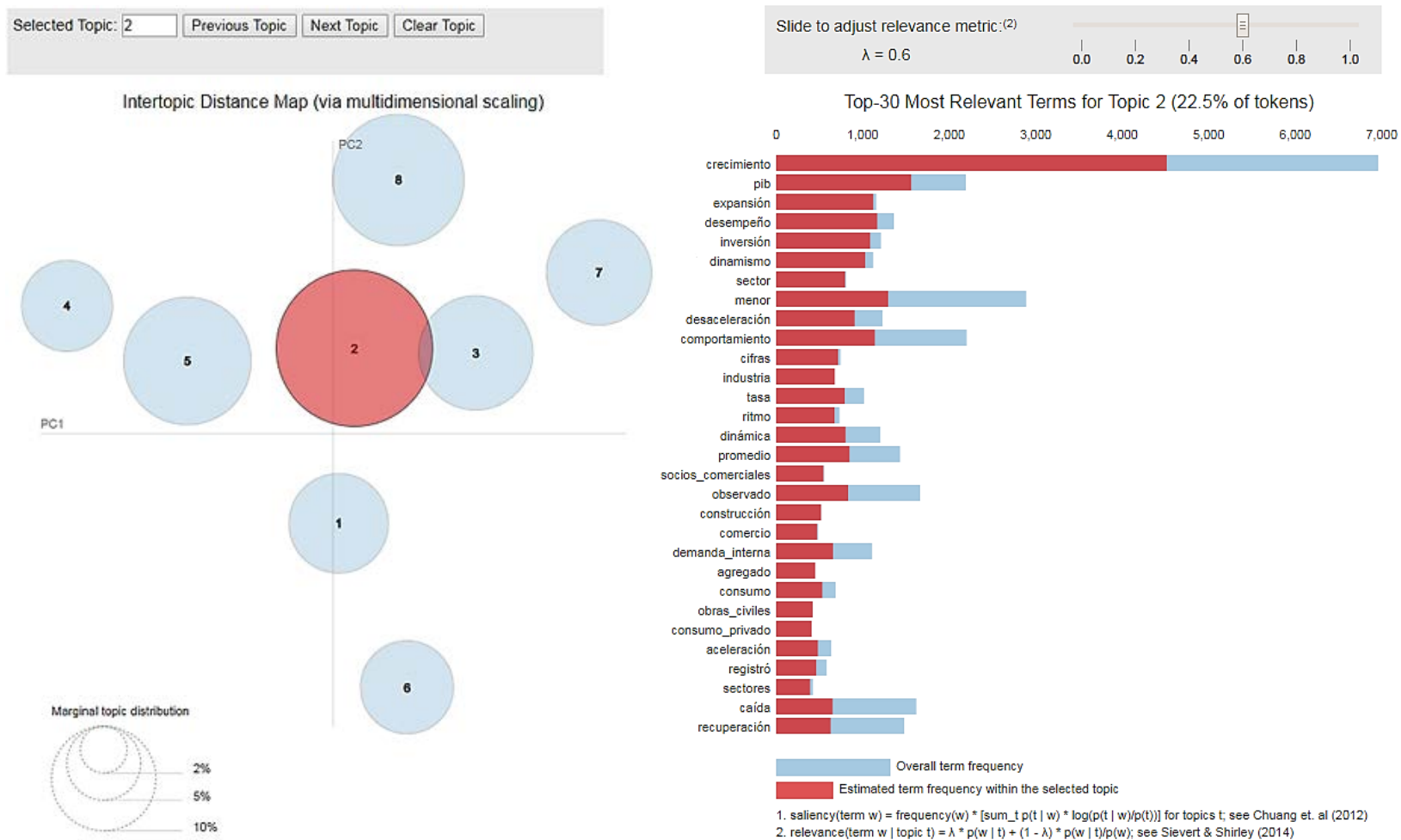
Para ilustrar el tipo de salidas que entrega el sistema LDAvis se usan como ejemplo los resultados obtenidos para el tema número 2 (que, como se explica abajo, hemos denominado “Demanda Interna y Sectores Económicos”), presentados en la Gráfica 2. Como se ve en el lado derecho de esa gráfica, las barras de color claro (azul) representan las frecuencias de cada término en todo el corpus, mientras que las barras de color oscuro (rojo) son las frecuencias de cada término en el tema específico. El orden de importancia de cada término (esto es, su posición de arriba hacia abajo) se establece a partir de una medida de *relevancia*. Esta medida depende de la ponderación λ que, intuitivamente, representa la importancia que se le da a la frecuencia del término en un tópico específico (relativa a su frecuencia en todo el corpus) versus la exclusividad del término para ese tópico (es decir, $\lambda = 0$ es equivalente a darle toda la relevancia a la exclusividad del término y ninguna a su frecuencia en el tópico). Más formalmente, λ determina el peso dado a la probabilidad del término w en el tópico k en relación con una medida de exclusividad conocida como *lift*¹².

El valor de λ puede ser ajustado a discreción del investigador. Para el presente documento se sigue a Sievert y Shirley (2014), que encuentran evidencia de que usar valores intermedios de λ mejora la interpretabilidad de los tópicos. Específicamente el

¹² El *lift* es la relación que hay entre la probabilidad de un término dentro de un tópico y su probabilidad marginal en el corpus. Esta medida disminuye la importancia de los términos globalmente frecuentes y le da más representatividad a los términos “exclusivos” del tópico. Sin embargo, clasificar los términos únicamente bajo esta medida podría afectar negativamente la interpretabilidad. Pueden encontrarse más detalles sobre el cálculo de la medida de relevancia en Sievert y Shirley (2014)

óptimo calculado por ellos es $\lambda = 0.6$ (y, por tanto, el valor usado para nuestros propósitos).

Por otro lado, para determinar la relevancia de cada tópico y la similitud entre tópicos, el sistema despliega un mapa de distancia intertemático (véase lado izquierdo de la Gráfica 2). Este muestra cada tópico como un círculo enumerado y en un plano de dos dimensiones, las cuales intentan capturar la variabilidad de los temas a partir de su reducción a dos factores (realizado a través del método estadístico de análisis de componentes principales). El tamaño de cada círculo representa la frecuencia relativa del tema en el corpus. El centro de cada uno es determinado con base en un algoritmo de escalamiento multidimensional aplicado a una matriz de distancia intertemática, de tal manera que la distancia entre los círculos sirva como medida de las diferencias entre tópicos: por ejemplo, los círculos que se interceptan representan similitudes en contenido (palabras probablemente similares). La medida que se usa por defecto para calcular la distancia intertópica es la divergencia Jensen-Shannon, la cual mide la similitud que hay entre dos distribuciones de probabilidad.



Gráfica 2. Visualización interactiva de los resultados del modelo LDA a través de la librería LDAvis. Elaboración propia.

2. RESULTADOS

Definición de etiquetas para los tópicos

Es importante recordar que el proceso computacional no produce etiquetas para cada tópico resultante. Es el investigador quien, opcionalmente, decide cómo etiquetar cada tema en función de la información obtenida. Como se ha explicado, los resultados del modelo LDA muestran cuáles son los términos relevantes presentes en cada tópico, y, basados en esa información, se pueden asignar nombres o etiquetas.

El Cuadro 1 muestra las primeras 10 palabras generadas para cada tópico, de acuerdo con los resultados de la medida de relevancia, junto con la participación de los tópicos en el corpus (minutas e IPM) a lo largo de todo el periodo analizado, y los nombres que se sugieren en el presente trabajo para etiquetarlos y hacer referencia a cada uno de ellos en lo que resta del mismo¹³.

Es importante señalar que, entre los resultados obtenidos, a partir del proceso explicado en la anterior sección, están la determinación del número de tópicos (ocho para este caso) y las distribuciones o proporciones de los mismos, periodo a periodo (de forma separada para las minutas y los informes). Estos dos resultados son los que se usan principalmente para el análisis que se presenta en esta sección.

Para facilitar la comparación entre minutas e IPM, se analizan los resultados para el periodo en que los dos han estado disponibles al público, esto es, desde junio de 2007 (cuando se comienzan a publicar las minutas).

¹³ Para asignar un nombre a cada tópico se han tenido en cuenta los términos más importantes asociados al mismo tanto a través de la medida de relevancia ($\lambda = 0.6$), como a través del ordenamiento que se genera cuando se considera únicamente la exclusividad de los términos ($\lambda = 0$). Sobre la medida de relevancia y la ponderación λ véase sección 1.3.

Tópico	2	8	5	3	7	1	6	4
Nombre	Demanda interna y sectores	Proyecciones macroeconómicas	Inflación, alimentos y regulados	Balanza comercial y materias primas	Contexto externo	Tasas de interés y política monetaria	Crédito	Inflación, meta y expectativas
Participación	22.5%	15.8%	14.9%	12.0%	10.2%	9.1%	7.9%	7.6%
Palabras	crecimiento	crecimiento	precios	exportaciones	estados_unidos	tasas_interés	crédito	inflación
	pib	pronóstico	inflación	términos_intercambio	zona_euro	caídas	cartera	meta
	expansión	esperado	alimentos	déficit_corriente	región	política_monetaria	empresas	expectativas_inflación
	desempeño	economía_colombiana	variación	importaciones	economías_emergentes	fed	recursos	inflación_básica
	inversión	inflación	inflación_consumidor	petróleo	china	tasas	financiamiento	situó
	dinamismo	proyecciones	transables	precio_petróleo	economías	tasa_interés	continúan	equipo_técnico
	sector	economía	regulados	déficit	américa_latina	mercado	precios_vivienda	indicador
	menor	previsto	aumentos	reducción	recuperación	actividad_real	deuda	banco_república
	desaceleración	crecimiento_económico	precios_alimentos	producción	primas_riesgo	política	transmisión	esperada
	comportamiento	senda	ajustes	demanda_externa	riesgo	dólar	préstamos	nivel

Cuadro 1. Primeras 10 palabras generadas por el software para cada tópico, ajustando $\lambda = 0.6$.

Minutas e Informes de Política Monetaria

Como parte de su propósito de transparencia y el esquema de comunicación con el público, el Banco de la República (BR) hace uso de varios tipos de documentos para divulgar y explicar su decisión de política monetaria y el análisis de la coyuntura económica y las proyecciones macroeconómicas en los que se basa dicha decisión. Uno de ellos son las *minutas* de las reuniones de política monetaria, esto es, las reseñas escritas de las consideraciones hechas por los miembros de la Junta Directiva del Banco de La República (JDBR) durante dichas reuniones. Las minutas se publicaban mensualmente y alrededor de dos semanas después de la correspondiente reunión. Desde 2018, su frecuencia se redujo como consecuencia de la reducción del número de reuniones de política monetaria a ocho por año. También a partir de ese año se redujo el lapso entre la reunión de política y la publicación de las minutas a alrededor de una semana¹⁴.

Otro documento fundamental en el proceso de comunicación con los mercados y el público general es el IPM. Se trata de un análisis y recopilación de información que sirven como insumo (entre otros documentos) para la deliberación de la JDBR. El IPM, durante el periodo analizado, fue publicado trimestralmente entre dos y tres semanas después de la correspondiente reunión de la Junta y contiene temas relativos al contexto externo, crecimiento interno, desarrollos recientes de la inflación y los pronósticos de mediano plazo¹⁵. Cabe resaltar que a diferencia de las minutas, que corresponden a la deliberación y análisis de la Junta, el IPM es elaborado por el equipo técnico de forma independiente.

¹⁴ A partir de octubre de 2019 las minutas se publican al siguiente día hábil de toda reunión ordinaria de Junta en la que se toma decisión de política monetaria (esto es, 8 veces al año, en los meses de enero, marzo, abril, junio, julio, septiembre, octubre y diciembre) y el IPM, de forma trimestral, al siguiente día hábil de las reuniones de Junta de los meses de enero, abril, julio y octubre. Como se mencionó en la introducción, nuestro periodo de análisis incluye información hasta diciembre de 2018.

¹⁵ Como se explica en la nota al pie anterior, desde octubre de 2019 el IPM se publica con un rezago de solo un día hábil con respecto a la correspondiente reunión de Junta.

Importancia de los temas según el ciclo de la tasa de política

Siguiendo un proceso de análisis similar al realizado por Hansen, McMahon y Tong (2018), inicialmente se ordenan los tópicos con base en un indicador del grado de ciclicidad con respecto a la tasa de política monetaria. Para construir dicho indicador, se separan los documentos teniendo en cuenta si se publican durante una fase de aumentos o una de disminuciones en la tasa de interés de política; luego, para cada tema se calcula su participación promedio en los documentos durante cada fase y se ordenan los temas en función de la diferencia entre el promedio durante los aumentos y el promedio durante las disminuciones. Los resultados se presentan en el Cuadro 2, tanto para las minutas como para los informes. Allí se incluye también el cálculo a partir de las medianas con el fin de presentar una medida que aisle datos atípicos; por ejemplo, el indicador a partir de promedios puede sesgarse hacia valores altos para un tópico que haya tenido una participación inusualmente alta en algún periodo específico.

Cuando la diferencia calculada para un tópico determinado es positiva (negativa), nos indica que ese tópico adquiere una mayor relevancia en las fases de aumento (disminución) de la tasa de política. Para estimar si la diferencia es significativa se realizó una prueba t de diferencia de medias (véase Cuadro 2). Como se puede ver en los resultados, excepto para el caso del tema de demanda interna, la importancia de los temas en el análisis del equipo técnico del Banco de la República (IPM) no parece cambiar dependiendo de las fases de aumento o disminución de la tasa de política. En contraste, la participación de algunos temas en el análisis para la deliberación de política monetaria de la Junta (minutas) sí se ve afectada por el ciclo de la tasa. La demanda interna y las expectativas de inflación parecen ser temas que adquieren mayor participación en la deliberación durante los periodos de disminuciones de la tasa de política, mientras que lo contrario sucede con los temas relacionados con las proyecciones macroeconómicas y el crédito.

Tópico	Promedio total	Aumentos (promedio)	Disminuciones (promedio)	Diferencia (promedio)	Diferencia (mediana)
MINUTAS					
2 Demanda interna y sectores	22.88%	21,10%	25,28%	-4,18%***	-4,51%
8 Proyecciones macroeconómicas	14.93%	15,66%	13,87%	1,79%***	1,39%
1 Tasas de interés y política monetaria	12.75%	12,88%	12,67%	0,21%	-0,60%
5 Inflación, alimentos y regulados	12.30%	13,23%	11,76%	1,47%	1,31%
4 Inflación, meta y expectativas	10.70%	8,72%	10,82%	-2,10%***	-1,83%
3 Balanza comercial y materias primas	10.09%	10,73%	10,37%	0,36%	0,78%
7 Contexto externo	10.00%	10,86%	9,39%	1,47%	1,59%
6 Crédito	6.37%	6,83%	5,84%	0,99%**	0,91%
INFORMES DE POLÍTICA MONETARIA					
2 Demanda interna y sectores	23.01%	21,53%	24,10%	-2,57%**	-3,32%
5 Inflación, alimentos y regulados	16.13%	16,69%	15,62%	1,07%	0.91%
8 Proyecciones macroeconómicas	15.67%	16,28%	15,36%	0,92%	1.05%
3 Balanza comercial y materias primas	12.89%	12,51%	12,68%	-0,17%	0.02%
7 Contexto externo	10.36%	10,70%	10,35%	0,35%	0.49%
6 Crédito	8.30%	8,14%	8,37%	-0,23%	-0,41%
1 Tasas de interés y política monetaria	7.30%	7,67%	7,08%	0,59%	0.56%
4 Inflación, meta y expectativas	6.34%	6,47%	6,44%	0,03%	0.02%

Cuadro 2. Cálculo del indicador de ciclicidad de los temas, con respecto al ciclo de la tasa de política.

El indicador mide la importancia del tópico dependiendo del ciclo de política monetaria (si es en aumentos o disminuciones de la tasa de política).

*** y ** denotan significancia al 1% y 5%, respectivamente, para una prueba *t* de diferencia de medias (bajo el supuesto de varianzas desconocidas y desiguales).

Análisis de la evolución en el tiempo de la importancia de los temas

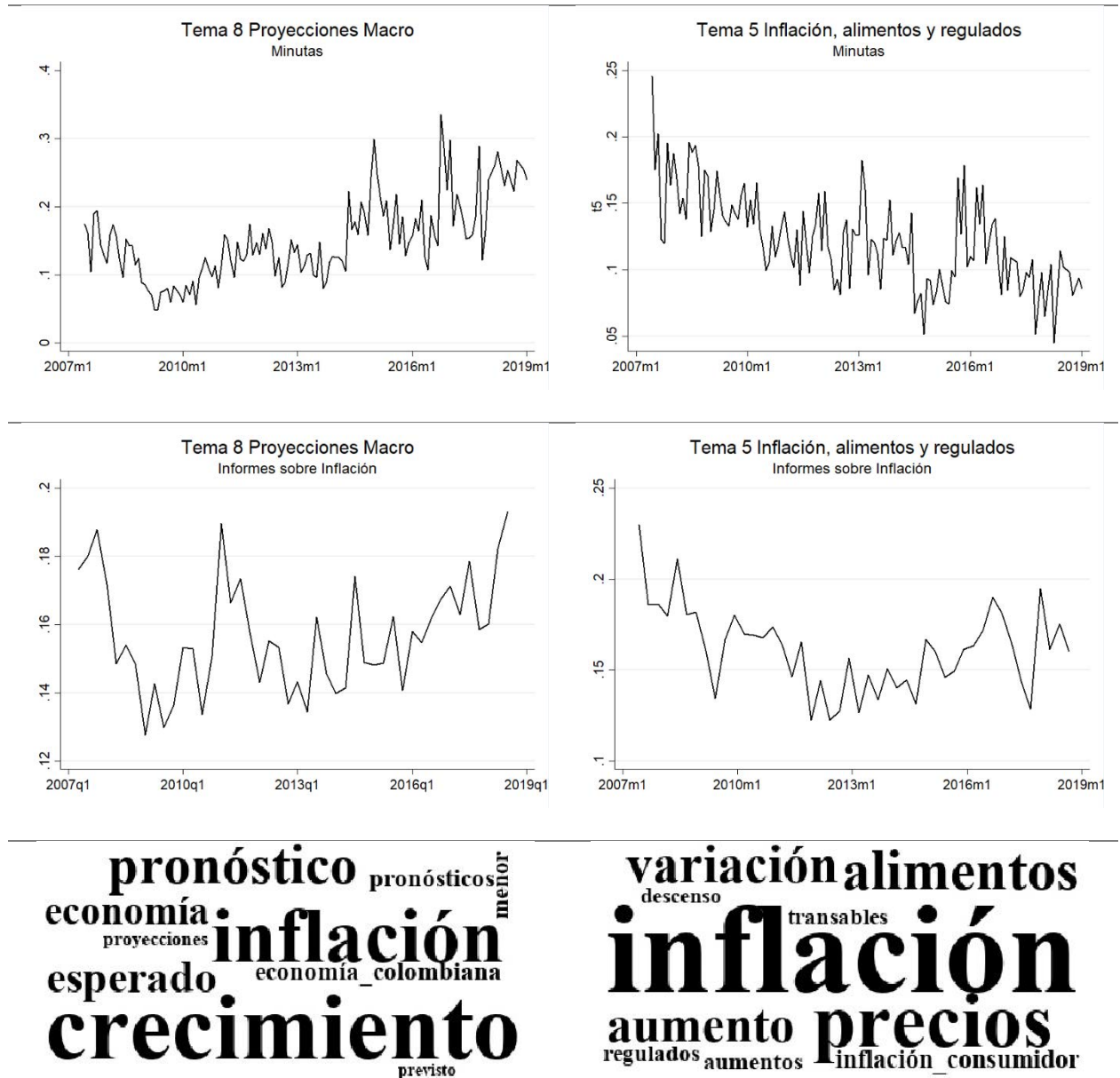
En el Cuadro 2 (arriba) y las gráficas que se incluyen en lo que sigue del documento (Gráficas 3-9), puede observarse que en varios temas hay diferencias significativas entre la participación que tienen en las minutas y la que tienen en los informes, y en la evolución de esas participaciones a través del tiempo. Los resultados que se presentan en lo que resta de esta sección, parecen indicar que la información narrativa de las minutas, entre 2007 y 2018, contenía un análisis con una mirada más prospectiva *-forward looking-*, que la de los IPM, durante el mismo periodo.

Lo anterior puede ser consecuencia del hecho de que los informes, de una extensión significativamente mayor que las minutas, ofrecían al público y los mercados un análisis de la dinámica reciente de la inflación y de la economía, incluyendo una parte importante de contenido retrospectivo. Por otro lado, debe también tomarse en cuenta que, en contraste con las minutas, una buena parte de la información que mira hacia adelante en el IPM es entregada en forma numérica (v.g. proyecciones macroeconómicas) y, por tanto, no es considerada dentro de la presente caracterización, basada en herramientas de análisis textual¹⁶.

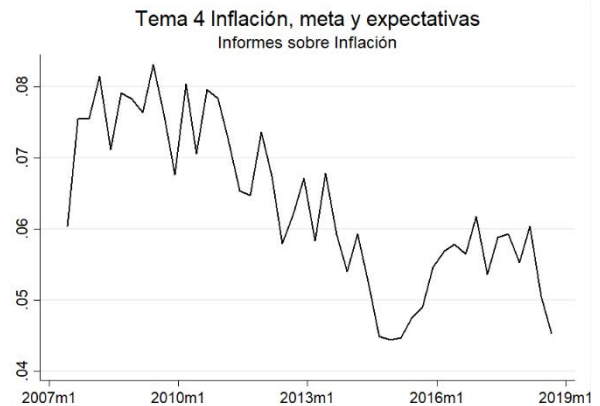
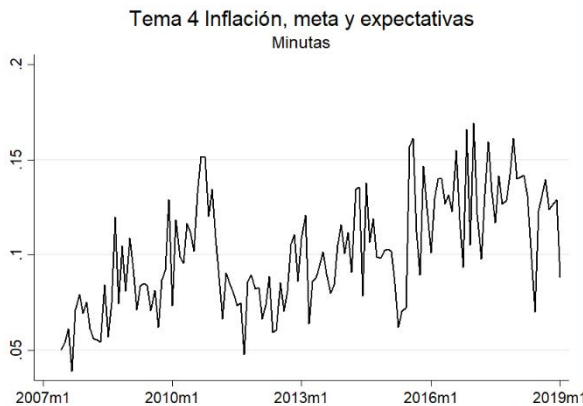
En la Gráfica 3 se observa que el tema de las proyecciones macroeconómicas (tópico 8) ha ganado relevancia desde 2009, principalmente en las minutas, alcanzando una participación en las mismas de alrededor de 25% durante 2018. En contraste, el tema de inflación de alimentos y regulados (tópico 5) ha perdido importancia en los dos tipos de documentos, aunque principalmente en las minutas. La evolución observada de la participación de estos dos temas (8 y 5) tanto en las minutas como en el IPM, podría ser consecuencia de una posible disminución de

¹⁶ A partir de octubre de 2019, el Banco de la República hizo modificaciones al contenido del IPM, las cuales aumentan el énfasis en el análisis de las proyecciones macroeconómicas. Dado lo anterior, es probable que al replicar en unos años un ejercicio similar al realizado en el presente documento, se encuentren menos diferencias en el carácter prospectivo de la información narrativa de los IPM con respecto a las minutas.

la relevancia de la indexación de precios a la inflación pasada en la explicación de la dinámica de la inflación en la economía colombiana.



Gráfica 3. Comparación participación del tema 8 “Proyecciones Macroeconómicas” y tema 5 “Inflación, alimentos y regulados”, en las minutas y los IPM.

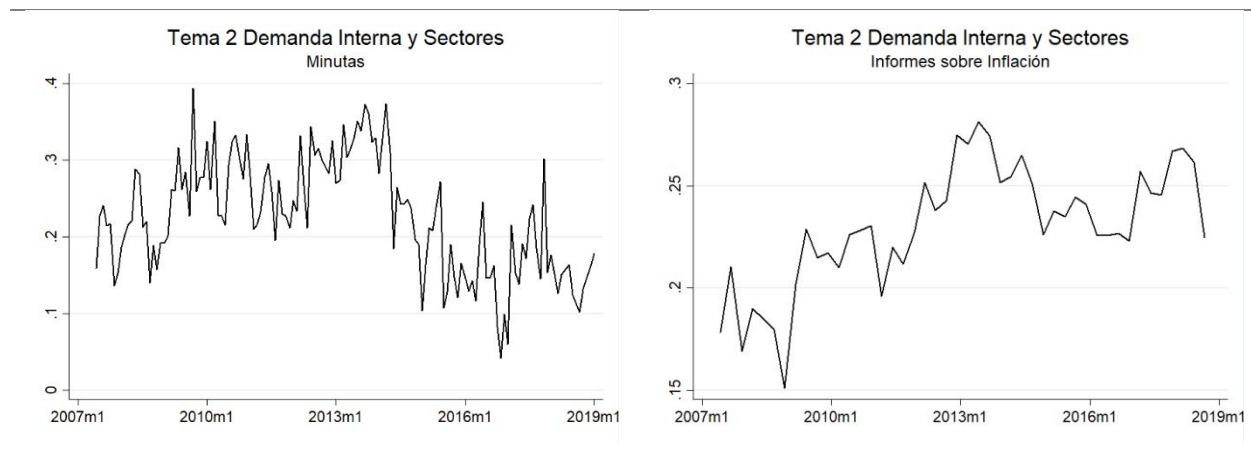


meta nivel promedio
 inflación básica expectativas
 esperada situó indicador
inflación
 expectativas_inflación

Gráfica 4. Comparación participación del tema 4 “Inflación, meta y expectativas”, en las minutas y los IPM.

Posiblemente relacionado también con las diferencias en la estructura del análisis que se hace en las minutas y el que se hace en los informes, se observa una tendencia creciente de la participación del tema de inflación, meta y expectativas (tópico 4) para el caso de las primeras, en contraste con una baja relevancia y una tendencia decreciente para el caso de la participación en los IPM (Gráfica 4).

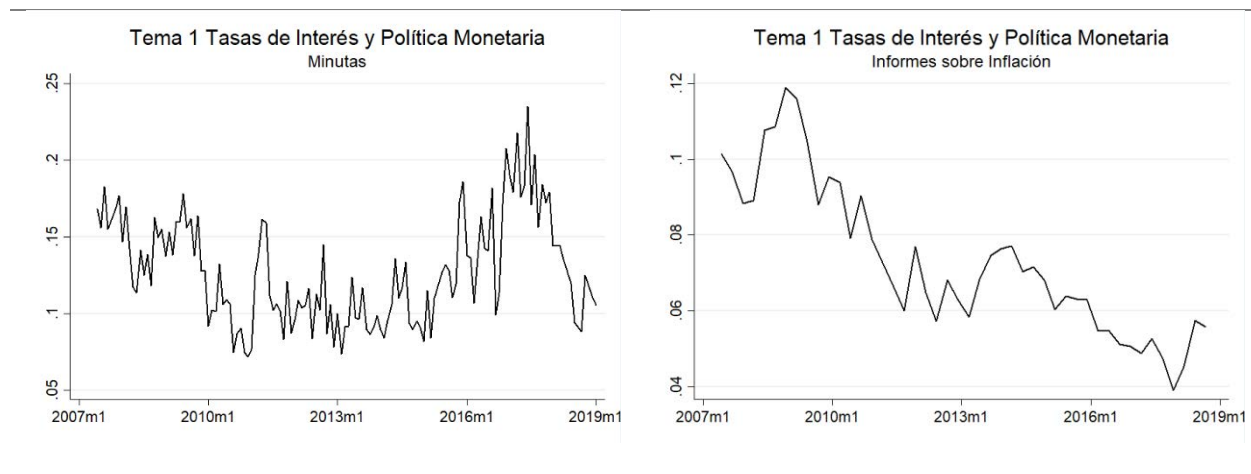
La demanda interna y los sectores económicos (tópico 2) han sido uno de los temas más relevantes a lo largo de todo el periodo analizado, tanto en las minutas como en los informes; sin embargo, desde 2014 ha perdido participación, en el caso de las minutas. En el caso de los IPM el tema ha incrementado su importancia a través del tiempo, especialmente con respecto a la que tenía al principio del periodo estudiado (Gráfica 5).



dinamismo menor
comportamiento expansión
crecimiento
desempeño pib inversión
desaceleración pib promedio

Gráfica 5. Comparación participación del tema 2 “Demanda interna y sectores”, en las minutas y los IPM.

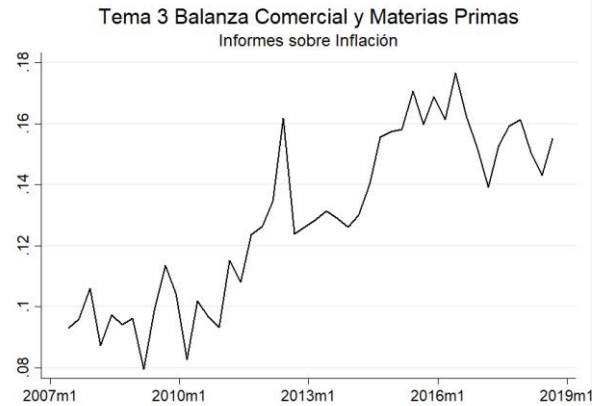
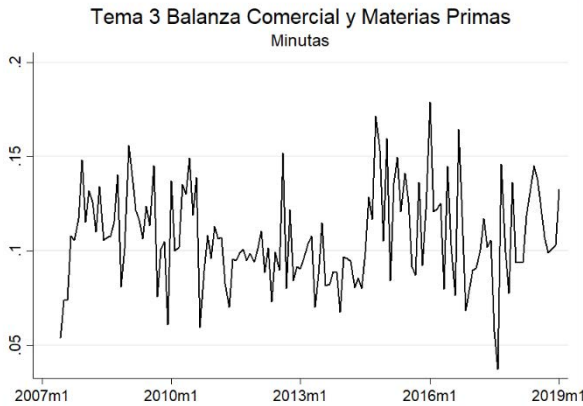
En la participación a largo plazo del tema 1, relacionado con tasas de interés y política monetaria, no se ve un cambio sustancial en las minutas durante el periodo analizado (Gráfica 6); sin embargo, entre comienzos de 2016 y hasta mediados de 2017 hubo un incremento importante. En las minutas de esos periodos, se evidencia la importancia que tenía para la Junta el posible efecto del proceso de “normalización” de la política monetaria que se estaba llevando a cabo en los Estados Unidos. En contraste, en los IPM el tópico 1 ha ido perdiendo importancia a través del tiempo.



aumento **tasas**
 caídas mercado **tasas_interés**
tasas_interés fed **reducción**
 política_monetaria incremento

Gráfica 6. Comparación participación del tema 1 “Tasas de interés y política monetaria”, en las minutas y los IPM.

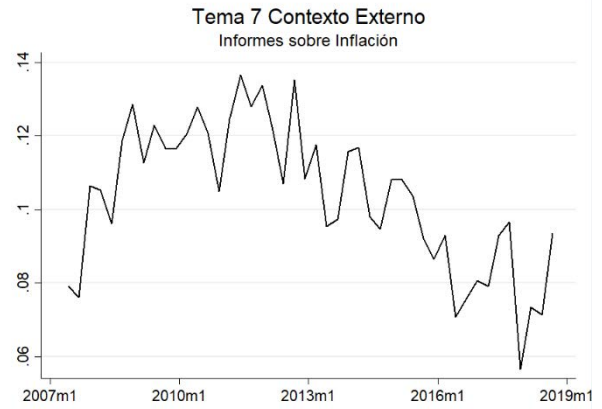
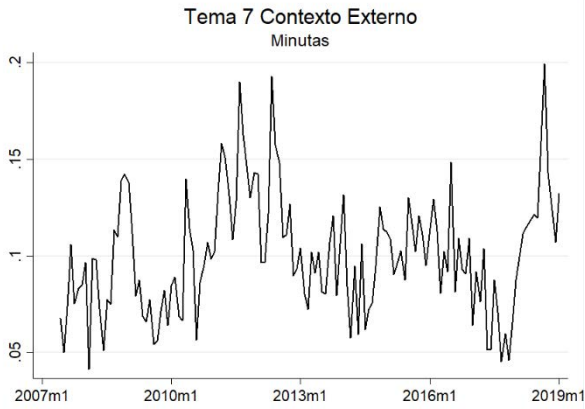
Con respecto al tema de la balanza comercial y materias primas (tópico 3), su importancia en las minutas fluctúa bastante en plazos muy cortos, pero se ha mantenido relativamente estable en plazos largos (alrededor del 10%). Su importancia en los IPM se ha incrementado desde niveles por debajo del 10% en 2007 a niveles por encima del 15% en 2018 (Gráfica 7). Se trata de un tema que usualmente está presente en el análisis dada la importancia de las materias primas, especialmente del petróleo, para el ingreso de la economía y la evolución de la tasa de cambio.



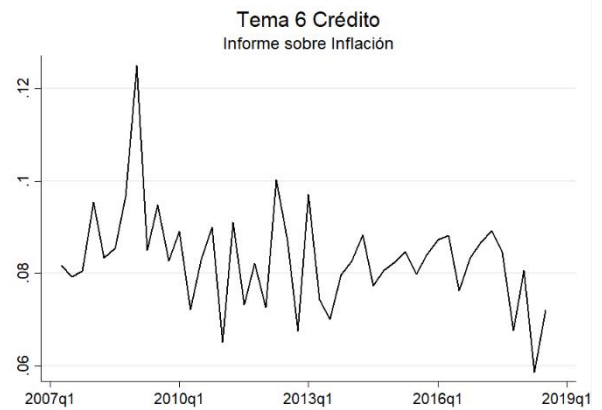
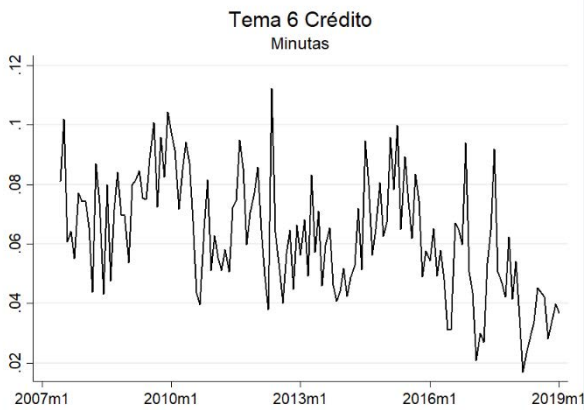
menor déficit corriente
 reducción — caída
exportaciones
 importaciones
 términos intercambio
 producción — observado petróleo

Gráfica 7. Comparación participación del tema 3 “Balanza comercial y materias primas”, en las minutas y los IPM.

La participación del tema “contexto externo” (tópico 7), tuvo un comportamiento fluctuante pero relativamente estable a plazos largos para el caso de las minutas. Se observan algunos incrementos significativos en momentos muy específicos, por ejemplo, entre agosto de 2011 y mayo de 2012. Durante este periodo, el equipo técnico del Banco redujo el pronóstico de crecimiento de los países desarrollados. La preocupación radicó en los temores de insolvencia fiscal de la Unión Europea y el potencial riesgo de una reducción del ingreso nacional generada por una demanda externa deprimida. A partir de 2013 el tema ha venido perdiendo relevancia en los IPM (Gráfica 8).



Gráfica 8. Comparación participación del tema 7 “Contexto externo”, en las minutas y los IPM.



Gráfica 9. Comparación participación del tema 6 “Crédito”, en las minutas y los IPM.

Finalmente, el tema del crédito (tópico 6) ha tenido una participación relativamente baja (comparado con la mayoría de los otros temas) tanto en las minutas como en los IPM (Gráfica 9).

3. CONCLUSIONES

A partir del uso de herramientas de lingüística computacional el presente trabajo analiza dos tipos esenciales de documentos en la comunicación del Banco de la República de Colombia, las minutas de las reuniones de política monetaria y los IPM (para el periodo marzo de 2007 a diciembre de 2018). Específicamente, se hace uso del método Latent Dirichlet Allocation (LDA), cuya finalidad es develar los patrones temáticos subyacentes que agrupan las palabras de un conjunto de documentos. En este método, cada documento puede verse como una distribución de probabilidad específica sobre los temas identificados.

Con base en el mencionado método, se encuentra que las minutas e IPM giran primordialmente en torno a ocho temas, siendo el más importante (en promedio a través del tiempo, tanto para las minutas como para los Informes) el que contiene términos principalmente relacionados con demanda interna y sectores económicos. Los otros siete temas son: proyecciones macroeconómicas; inflación, alimentos y regulados; balanza comercial y materias primas; contexto externo; tasa de interés y política monetaria; crédito; inflación, meta y expectativas.

Con respecto a la importancia de cada tema dentro de los documentos y la evolución de esa relevancia en el tiempo, se encuentran diferencias significativas entre las minutas y los informes. Por ejemplo, después de 2012 se observa una tendencia creciente de la participación del tema de inflación, meta y expectativas para el caso de las minutas, en contraste con una baja relevancia, a través de todo el periodo analizado, y una reducción de la misma durante buena

parte de ese periodo, para el caso de los informes. Los resultados muestran, sin embargo, que existen algunas similitudes en la evolución de las distribuciones de temas de minutas e informes. En particular, cabe destacar la disminución de la participación del tema de inflación de alimentos y regulados y el aumento de esta para el caso del tema de las proyecciones macroeconómicas.

BIBLIOGRAFÍA

- Arango, L. E., Pantoja, J., & Velásquez, C. (2017). Effects of central bank's communications in Colombia. A content analysis. Borradores de Economía.
- Bancolombia (2019). Las comunicaciones del Banco de la República y sus implicaciones de política monetaria. Reporte Especial, Investigaciones Económicas, Grupo Bancolombia. Julio
- Bholat, D., Hansen, S., Santos, P., & Schonhardt-Bailey, C. (2015). Text mining for central banks: handbook. Centre for Central Banking Studies(33), 1-19.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993-1022.
- Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn & TensorFlow. O'Reilly.
- Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. The Quarterly Journal of Economics, 801-870.
- Hansen, S., McMahon, M., & Tong, M. (2018). The Long-Run Information Effect of Central Bank Communication. Staff Working Paper(777).

- Hoffman, M. D., Blei, D., & Bach, F. (2010). Online learning for Latent Dirichlet Allocation. NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems, 856-864.
- Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the Space of Topic Coherence Measures. Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, 399-408.
- Sarkar, D. (2016). Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data. apress.
- Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, 63-70.
- Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. IEEE International Conference on Data Science and Advanced Analytics (DSAA), 165-174.
- Taborda, R. (2015). Procedural transparency in Latin American central banks under inflation targeting schemes. A text analysis of the minutes of the Boards of Directors. Ensayos sobre Política Económica, 33(76), 76-92.

